Improved Multivariable Statistical Process Control (MSPC) for Chemical Process Fault Detection and Diagnosis (PFDD) – Cross-Variable Correlation Approach

Lam Hon Loong and Kamarul Asri Ibrahim

Faculty of Chemical and Natural Resources Engineering
Universiti Teknologi Malaysia, 81310 Skudai, Johor, Malaysia

Email: honloong@tm.net.my

ABSTRACT

Chemical process is inclined to be large-scale, continuous, integrated and linkable around the whole plant. It has put forward tough requirements on desired quality, high yield/production, low consumption, environmentally friendly and safe operation. Thus accurate process fault detection and diagnosis (PFDD) system at early stage of the process are very important to modern chemical plant in achieving the above vision. To ensure that the product is obtained in the acceptable limit ranges, some monitoring and control methods have to be applied. This paper focuses on the application of Multivariate Statistical Process Control (MSPC) as the fault detection and diagnosis tool. This is because MSPC method can easily monitor large volumes of highly correlated data, and the process information is clearly presented in the analysis chart. The paper explains how the cross-variable correlation coefficient and the multivariate control chart are developed. When a fault(s) is detected, the cause of the fault is diagnosed by means of contribution charts. The result, which is based on this new approach perform better compared to those based on conventional MSPC analysis.

Keywords: multivariate statistical process control (MSPC), process fault detection and diagnosis (PFDD), T^2-statistics, cross-variables correlation coefficient

1.0 Introduction

The main objective of a PFDD system is to define fault precisely in advance by making comparisons, which indicate whether or not a fault occurs. After detection of a fault occurrence, it's followed by determination of where the fault came up (cause), thus fault can be identified. In others way we can say that PFDD is to able to decide, given current process conditions, whether the process is normal or abnormal, and if abnormal, whether the cause is a known fault or an unknown fault; if it is a known fault, identify the source of the fault.

Currently, in many chemical processes, they are becoming increasingly measurement rich. Large volume of highly correlated data is always recorded. If an appropriate tool is applied, this large volume of data can be very useful for process monitoring (Lam and Kamarul, 2002). Multivariate Statistical Process Control (MSPC) is one of the proposed tools to extract information from the data by carrying out data reduction without losing the previous information. The typical univariate approach method such as Statistical Process Control (SPC) is not applicable in chemical process, which is multivariate in nature. Many industrial processes involve a set of input variables, process variables and output variables, which are highly correlated. If one of the variables changes, it will affect other correlated variables. Ignoring the influence or cross-correlation between the variables can lead to faulty interpretation about the control process. For an example, a small change in univariate data may not be significant enough to raise an alarm and be considered as a fault. However, if taken as whole, many "small" correlated problems may signify a major change in the plant.

One of the advantages of MSPC is that the method could reduce the complexity of online performance monitoring with its ability to detect process or product abnormalities that are difficult to notice. MSPC uses statistical projection techniques of Principal Component Analysis (PCA) to reduce the dimensionality of the problem (Lennox et al, 2000). These new variables explain the maximum amount of variability in the data and make sure that the information of the original data will not be lost.

After the highly correlated data has been reduced, Multivariate Control Chart (T^2-control chart or cross-variable correlation control chart) is plotted to monitor process performance. The function of this control chart is to compare with the current state of the process with "Normal Operating Condition (NOC)". The "NOC" condition exists when the process or product variables remain close to their desired values. In contrast, the "Out of Control (OC)" occurs when fault appears in the process. While, fault (s) is detected, the cause of the problem is diagnosed by using contribution chart.

2.0 Process Modeling and Data Generation

Data collection is the most important part in PFDD. In this work, the data is obtained from a model simulation. The following section discusses the transformation of data to a multivariate statistical analysis. A typical cooled tubular reactor system for Ethylene Oxide production is selected and studied. The model is developed, based on the Westerterp and Ptasinki Model (1984). Figure 2 shows the process and all the variables involved.

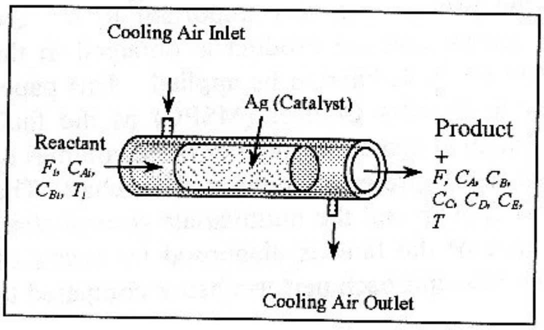

Figure 2: Schematic Diagram of Cooled Tubular Reactor System for EO Production

Where,
C_A = Concentration of Ethylene (mol/m^3)
C_B = Concentration of Oxigen (mol/m^3)
C_C = Concentration of Ethylene Oxide (mol/m^3)
C_D = Concentration of Carbon Dioxide (mol/m^3)
C_E = Concentration of Water (mol/m^3)
T = Reactor Temperature (K)
F = Reactants/ Products flowrate (m^3/s)
Tc = Coolant Temperature (K)
Fc = Coolant flowrate(m^3/s)

The state equations for the reaction process were derived based on first principals equation and formed ordinary differential (ODE) equation. These ODE state equations were solved by 4th order Runge-Kutta method. The MATLAB® software was used to graphically visualize the immediate dynamic response of the system (Ha, 1999).

Based on this reactor modeling, two sets of process operating data are generated. To generate NOC data, some noises were imbedded into the simulator. The noises considered are the small random change in the input flow rate and inlet coolant flow rate. On the other hand, to generate OC data, some significant, and insignificant fault appearances and also multiple fault appearances were purposely added into the process simulation. This OC data was compared with the NOC data in the analysis part later. The NOC and OC data information was summarized in **Table 1**. After this, the data was passed to multivariate statistical analysis, which would implement principal component analysis (PCA), T^2-control chart and contribution chart.

3.0 Get The Statistical Relationship Between the Collected Data via Eigenvector-Eigenvalue Decomposition Principal Component Analysis, (PCA).

Eigenvector-eigenvalue Decomposition PCA was applied in this research to form a PCA model for the reactor operating condition. For adaptive PCA, the PCA model is continuously updated where variables' sample mean, and covariance matrix are thus adapted to the changing conditions. A data-sampling scheme is proposed. Every sample contained 50 data and is updated for every 5-interval lapse. At the same time, new observed sample's mean and covariance matrix is recalculated and is considered to construct a continuously updated adaptive PCA.

Table 1: Fault Appearance Information in OC Data

Fault Appearance (i^{th} Data)	Normal Operating Condition	Fault Description	Fault Appearance (i^{th} Data)	Normal Operating Condition	Fault Description
250-280	Tc = 298.3	350	3850-3880	Fc = 0.051	0.035
750-780		311.65	4250-4280		0.049
12501-280	F = 0.221	0.19		**Multiple Fault Description**	
1750-1780		0.21			
2250-2280	Ca = 1.122	0.75	4550-4580	Tc = 350	
2750-2780		0.83		F = 0.19	
3250-3280	T = 300.15	315	4850-4880	T = 315	
3550-3580		304.23		Fc = 0.049	

The Second step is to standardize the input sample data and construct correlation matrix for all related variables. The data sample obtained should be standardized before further statistical analysis because the data was measured on the scales with widely differing ranges and the measurement units differ.

The next step is to calculate the eigenvalue, λ and eigenvector, v for the standardized samples, which are given in the following equation:

$$|X - \lambda I| = 0 \qquad \textbf{(Eq.1)}$$

$$Xv - \lambda v = 0 \qquad \textbf{(Eq. 2)}$$

Here, **X** is a standardized square matrix and ***I*** is the identity matrix of **X**.

Arrange the eigenvalue from the largest to the smallest then the accompanied eigenvectors matrix is determined.

The definition of Principal Components indicates that, a n new variables can be generated from the original n variables from data matrix X. Each new variables is a linear combination of the original variables such as:

$$p_{ci} = v_{1i}x_1 + v_{2i}x_2 + \cdots + v_{ni}x_n \qquad \textbf{(Eq. 3)}$$

From the linear equation given in Eq. 4, obviously, each new variable called principal component, Pc contain every information and relation between all original variables in the original matrix X (m measurements and n variables). The new Principal Component matrix is define by:

$$\mathbf{Pc = XV} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix} \begin{bmatrix} v_{11} & v_{12} & \cdots & v_{1n} \\ v_{21} & v_{22} & \cdots & v_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ v_{n1} & v_{m2} & \cdots & v_{nn} \end{bmatrix} \qquad \textbf{(Eq. 4)}$$

then,

$$\mathbf{Pc = XV} = \sum_{i=1}^{m}\sum_{j=1}^{n} x_{ij}v_{jk}; k = 1,2,\cdots,n. \qquad \textbf{(Eq.5)}$$

In this study case, only the eigenvalue for top three principals components (PC) that are greater than one are used. These covered and explained the large region of the total variation, thus, PC_1, PC_2 and PC_3 were selected to represent the original process behavior. Afterwards, the process states variables can be reduced to three, as a replacement for 6 original variables. Figure 3 shows that the original variables are replaced by three new variables called principals component. Simultaneously, the sample data's matrix was also reduced from 6 x 50 matrix to 3 x 50 matrix.

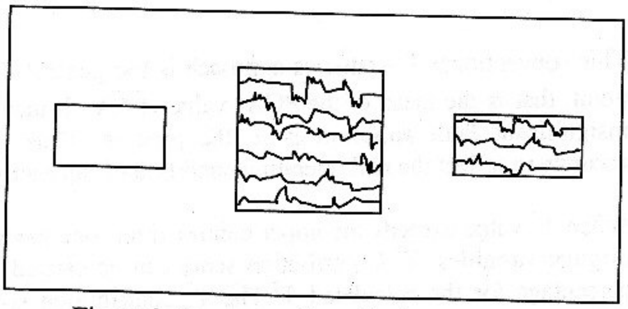

Figure 3: Parameter Complexity Reduction By Principals Component Analysis.

4.0 Conventional Process Fault Detection and Diagnosis

For monitoring a process with MSPC method, statistical control limit is set by T^2-statistics, after a PCA model is developed in section 3.0. A measure of the variation within the PCA model is given by Hotelling's T^2 statistics. T^2 is the sum of normalized squared score.

On the assumption of multivariate normality, it shows that distribution of the conventional T^2-statistics is given by Mason and Young, 1995:

$$T^2 = n\left(\overline{PCA}_{obs} - \overline{PCA}_{NOC}\right)^T \Sigma^{-1}\left(\overline{PCA}_{obs} - \overline{PCA}_{NOC}\right) \quad \textbf{(Eq. 6)}$$

and, the critical distance, C^2 (control limit) line is given by,

$$c^2 = \frac{p(n+1)(n-1)}{n(n-p)} F_{p,n-p}(\alpha) \quad \textbf{(Eq. 7)}$$

Where,

n = number of observation = 50; p = number of variable = 3

α = level of significance = 99%

F = F-distribution and $F_{3,\,57\,(0.01)} = 0.97$

A critical distance from the F-distribution is used to determine the upper control limit (UCL) for the T^2-control chart. Also, it defines the locus of the ellipse that contained the NOC data. A signal will be detected in the control chart when T^2 value of an incoming observation exceeds the value in Equation 7; or in the same way, if $T^2 > ULC$, a fault is detected. The T^2-control chart in Figure 4 presents the result of the conventional T^2-statistics analysis. There are only 8 out of 12 faults are detected.

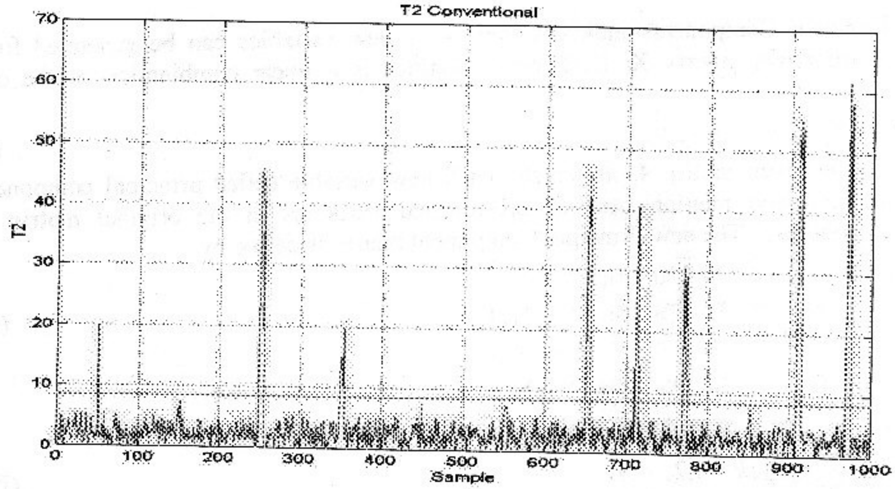

Figure 4: Conventional T^2 Statistics Control Chart.

This conventional T^2-statistics approach is too general. It summarized all the variables information in one point, that is the mean of the PCAs value, $\overline{\mathbf{PCA}}$. It may not be sensitive enough to analyze and detect the insignificant fault appearances in the process. Thus, the T^2-control chart result is not sensitive and accurate to reflect the real operating condition of the continuous process.

When T^2 value exceeds the upper control limit, one has to revert the determined T^2 value to the particular original variables. T^2 Contribution score can be carried out in the next step. Examine their contribution percentage for the calculated T^2.The T^2 contribution score for particular variable x_j, c_j is given by the following equation:

$$c_j = (x-\bar{x})_j^T (\mathbf{S})_j^{-1} (x-\bar{x})_j \quad \textbf{(Eq. 8)}$$

A large value of calculated T^2 contribution score associated with a particular variable, implying a corresponding large contribution to push forward the T^2 point out of the *ULC*. So, the variable, which contributed the highest score in the contribution chart, is the variable that caused the process to be out of control. The result of contribution plot was presented in later part, and it was compared to the result achieved by the modified method.

5.0 Cross-Variables Correlation Coefficient Between Variables Based on PCA Calculation.

The goal of this section is finding a correlation coefficient between variables. Thus, the advancement of this research is involved finding new relations between variables, where measuring such relations in the most straightforward manner. Actually, a lot of procedures that are described in this paper can be interpreted in terms of evaluating various kinds of inter-variable relations. A mathematical calculation based on PCA approach was proposed here.

Principal Component, P_C can be built up using several methods. As an example, the previous part presented the P_C calculation via Eigenvalue-Eigenvector Decomposition (EED). There is another powerful technique that can be used to analyze the equation or matrices that are either singular or numerically very close to singular. This technique is known as Singular Value Decomposition (SVD)(Milan & Ivan, 1991).

Let $\mathbf{X}$ is an $(m \times n)$ matrix as used in the previous section. It can be written as the product of an $(m \times n)$ column-orthogonal matrix $\mathbf{U}$, a diagonal matrix $\mathbf{L}^{1/2}$ with positive element, and the transpose of and $(n \times n)$ orthogonal matrix $\mathbf{V}$ (Kamarul, 1996), which is define as:

$$\mathbf{X} = \mathbf{U}\mathbf{L}^{1/2}\mathbf{V}^T \qquad \textbf{(Eq.9)}$$

Where,
$\mathbf{U}$ = Eigenvector matrix of $\mathbf{X}\mathbf{X}^T$,
$\mathbf{V}$ = Eigenvector matrix of $\mathbf{X}^T\mathbf{X}$,
$\mathbf{L}^{1/2}$ = Singular value,
= Diagonal matrix of positive square roots of the eigenvalue of $\mathbf{X}^T\mathbf{X}$,
$= \left|\sqrt{\lambda}\right|$.

The matrices $\mathbf{U}$, $\mathbf{V}$ and $\mathbf{L}^{1/2}$ have the following properties (Nash and Lefkovitch, 1976):

$$\mathbf{U}\mathbf{U}^T = \mathbf{U}^T\mathbf{U} = \mathbf{I} \qquad \textbf{(Eq.10)}$$

$$\mathbf{V}\mathbf{V}^T = \mathbf{V}^T\mathbf{V} = \mathbf{I} \qquad \textbf{(Eq.11)}$$

$$(\mathbf{L}^{1/2})(\mathbf{L}^{1/2})^T = (\mathbf{L}^{1/2})^T(\mathbf{L}^{1/2}) = \lambda \qquad \textbf{(Eq.12)}$$

The previous section introduced the mathematical interpretation of the new variables via PCA. The Principal Components are just linear combination of the original variables, which explain progressively smaller portions of total sample variance. Now, the following part shown, how there new variables information (eigenvector and eigenvalue) can be utilized to correlate between the quality-interested variable, $\mathbf{x}_k$ and the input variables, $\mathbf{x}_i$. Where,

$$\mathbf{x}_i = C_{ik}\mathbf{x}_k \qquad \textbf{(Eq.13)}$$

Correlation coefficient between $\mathbf{x}_i$ and $\mathbf{x}_k$ can be determined by following equation (Johnson and Wirchen, 1996):

$$corr(\mathbf{x}_i, \mathbf{x}_k) = \frac{\mathrm{cov}(\mathbf{x}_i, \mathbf{x}_k)}{\sqrt{\mathrm{var}\,\mathbf{x}_i}\sqrt{\mathrm{var}\,\mathbf{x}_k}} \qquad \textbf{(Eq.14)}$$

Since $\mathbf{X}$ is in the standardized matrix form, var($\mathbf{x}_i$) and var($\mathbf{x}_k$) = 1(Kamarul, 1996). So that the correlation coefficient from Eq.13 could be simplified and rearranged become:

$$C_{ik} = corr\,(\mathbf{x}_i, \mathbf{x}_k) = cov\,(\mathbf{x}_i, \mathbf{x}_k) = \mathbf{x}_i^T\mathbf{x}_k \qquad \textbf{(Eq.15)}$$

Multiplying both side of Eq.4 with $\mathbf{V}^T$,

$$\mathbf{X} = \mathbf{Pc}\mathbf{V}^T \qquad \textbf{(Eq.16)}$$

From the above equation (Eq.16), then,

$\mathbf{x_i} = \mathbf{Pcv^T}_i$ **(Eq.17)**

and, $\mathbf{x_k} = \mathbf{Pcv^T}_k$ **(Eq.18)**

To make these mathematic equations easy, let $\mathbf{V^T} = \tilde{\mathbf{V}} = [\tilde{\mathbf{v}}_1 \quad \tilde{\mathbf{v}}_2 \quad \cdots \quad \tilde{\mathbf{v}}_n]$, so the Eq.17 and Eq.18 can modified and become:

$\mathbf{x_i} = \mathbf{Pc}\,\tilde{\mathbf{v}}_i$ **(Eq.19)**

and, $\mathbf{x_k} = \mathbf{Pc}\,\tilde{\mathbf{v}}_k$ **(Eq.20)**

Let substitute the both equations above (Eq.19 and Eq.20) into equation Eq.15, then the correlation coefficient, C_{ik} is therefore:

$$C_{ik} = cov\,(\mathbf{x_i}, \mathbf{x_k}) = \mathbf{x_i}^T\mathbf{x_k}$$
$$= (\mathbf{Pc}\,\tilde{\mathbf{v}}_i)^T (\mathbf{Pc}\,\tilde{\mathbf{v}}_k)$$
$$= \tilde{\mathbf{v}}_i^{\,T}\,\mathbf{Pc^T Pc}\,\tilde{\mathbf{v}}_k \qquad \textbf{(Eq.21)}$$

Now, lets combine the PCA equations that derived via eigenvector-eigenvalue decomposition in Eq.4 and SVD decomposition in Eq.9. The principal component, Pc of the data matrix X can be rearranged:

$$\mathbf{Pc} = \mathbf{XV} = \mathbf{UL^{1/2}} \qquad \textbf{(Eq.22)}$$

To work up the Eq.18, we have to utilize Eq.19. From Eq.19, $Pc = UL^{1/2}$ then,

$$\mathbf{Pc^T Pc} = (\mathbf{UL^{1/2}})^T (\mathbf{UL^{1/2}})$$
$$= (\mathbf{L^{1/2}})^T \mathbf{L^{1/2} U^T U} \qquad \textbf{(Eq.23)}$$

According to Eq.8 and Eq.10, Eq.20 becomes,

$$\mathbf{Pc^T Pc} = \lambda \qquad \textbf{(Eq.24)}$$

Insert Eq.24 into Eq.21. Finally, the correlation coefficient between variable-*i* and variable-*k* simplified as:

$$C_{ik} = cov\,(\mathbf{x_i}, \mathbf{x_k})$$
$$= \tilde{\mathbf{v}}_i^{\,T}\,\lambda\,\tilde{\mathbf{v}}_k \qquad \textbf{(Eq.25)}$$

If *j* variables are involved, the correlation between the variable $\mathbf{x_i}$ and $\mathbf{x_k}$ based on the above equation can be rewrite in form of the transformed variables from PCA as:

$$C_{ik} = \sum_{j=1}^{n} v_{ij} v_{kj} \lambda_j \qquad \textbf{(Eq.26)}$$

So, the quality-interested variable, Concentration of Ethylene Oxide, Cc, can be related with others input variables in following equation:

$$Cc_k = C_{ik}\mathbf{x_k} \qquad \textbf{(Eq.27)}$$

Where, $\mathbf{x_k}$ are input data matrices such as C_A, T, F, Fc and Tc. The means of this determined Cc_k then is compared to the Cc_k means obtained from the NOC data. The general T^2-statistics equation from Eq.6 is applied and modified become:

$$T^2 = n\left(\overline{Cc}_{k\,obs} - \overline{Cc}_{k\,NOC}\right)^T \Sigma^{-1}\left(\overline{Cc}_{k\,obs} - \overline{Cc}_{k\,NOC}\right) \qquad \textbf{(Eq.28)}$$

The T^2-control chart in Figure 5 presents the result of the Cross-variable Correlation Coefficient T^2-statistics. This control chart succeeded in pointing out 10 out of 12 faults that had been put into the simulator. Compare to the conventional control chart result, conventional T^2-control chart is only able to detect significant changes in operation and high impact faults. An insignificant fault is unlikely to be detected in the conventional T^2-control chart. Thus, in term of sensitivity, cross-variable correlation coefficient control chart yields a better result.

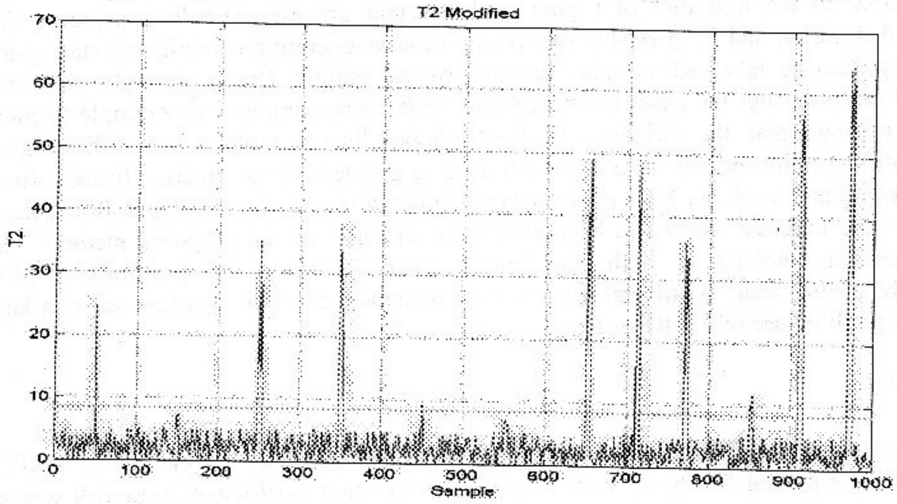

Figure 5: Modified T^2 Statistics Control Chart.

When T^2 value exceeded the upper control limit, again we have to revert the determined Cc_k value to the particular original variables $\mathbf{x_k}$. The contribution score equals to the Cc_k. Since our input and out matrix is in standardised form, the limit range for determined Cc_k is in the ranges of ± 3 standard derivations and hence, the control limit for the contribution chart, can be assumed as 3, where,

$$\text{Contribution Score Limit} = |C_{ik}\mathbf{x_k}| = 3. \qquad \textbf{(Eq.29)}$$

Fig. 6 and Fig.7 show the comparison between fault diagnosis results by contribution chart based on deference method. Fig. 6 is the result obtained from the conventional approach, which discussed in the previous part and Fig.7 is the diagnosis result from the cross-variable correlation coefficient approach. Basically, there are 8 out of 10 faults detected in Fig. 4 and 10 out of 12 faults detected in Fig.5. We tried to use contribution chart to find out the cause of these problem.

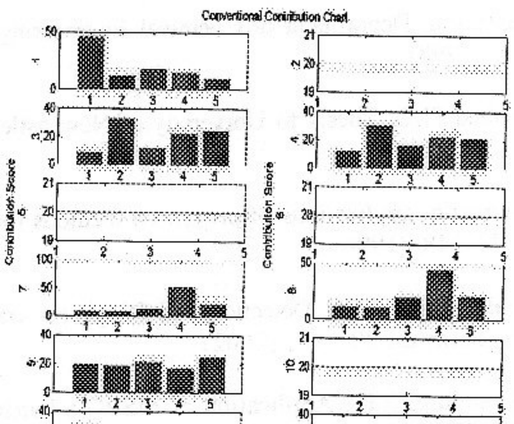

Fig. 6: Conventional Contribution Chart

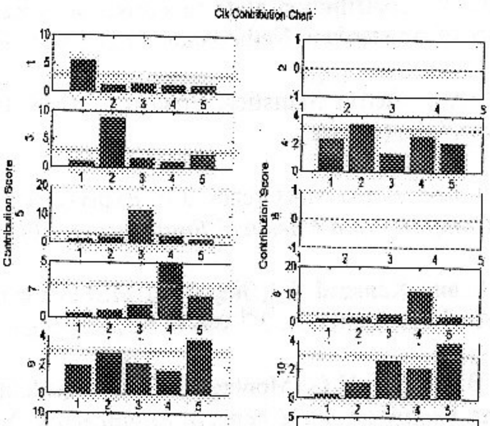

Fig. 7: C_{ik} Contribution Chart

Figure 6 and show contributions to the T^2 statistics for the reactor process. When the signal was detected from the T^2 control chart, the observed sample is automatically transferred to fault diagnosis stage. The first process variable in the chart was Inlet coolant temperature, Tc. It was followed by Reactant Flow rate, F, Concentration of Ethylene, C_A, Reactor temperature, T and coolant flow rate, Fc. The highest contribution variable is considered as a root of the problem at that particular period. For example, the problem in the process stayed within samples 50 to 56. Zooming in on this period revealed the root of the fault that occurred. In the first plot of Figure 6 shows that process variable 1 achieves the highest contribution at this period and *Tc* probably caused the process-operating condition to start going wrong. A similar strategy can be followed for other detected period sample. Disappointingly, this contribution chart still could not point out the exact multiple faults in last two samples, for example in plot-11. The plot-11 in Fig. 6 shown that, the variable-1 is slightly higher than variable-2. It is difficult to verify the cause of the fault and point out the variable which making problem to the process. In fact, the fault was contributed by combination of the high inlet coolant temperature and the low input flow rate, which is listed in table 1 and they are succeed to be pointing out in Fig.7 by the proposed method. The similar condition happened in plot-12 for both contribution charts. However, the performance of the fault detection with the control chart is still unsatisfied. This remains a research problem and is a key element in improving the performance of PFDD system.

CONCLUSION

An approach to develop an efficient fault detection and diagnosis scheme by the use of MSPC method was presented. The improved MSPC by Cross-variable Correlation Coefficient Approach was suggested to tackle the dynamic operating condition for multiple-reactor. The results have confirmed the efficiency of the proposed approach. The sensitivity of this proposed approach is better than those achieved by the conventional method. The monitoring of the conventional T^2-control chart is not always accurate, particularly if the fault is insignificant. The advantage of the proposed approach is that the multiple faults can be easily diagnosed. As a result, Cross-variable Correlation Coefficient method is more sensitive then conventional method, even though a small undesirable change in the process occurs.

ACKNOWLEDGMENTS

National Science Fellowship (NSF) scholarship is gratefully acknowledged.

REFERENCES

Ha, T.C., Feedback control of cooled tubular reactor for multiple reaction. Universiti Teknologi Malaysia: Undergraduate Thesis. (1999).

Johan, A. W. Contribution plots in statistical process monitoring, Department of Chemical Engineering, University of Amsterdam, Netherlands: PhD Thesis Proposal. (1998)

Kamarul, A.I. Active Statistical Process Control. Unpublished PhD. Thesis to University of Newcastle, Upon Tyne, UK. (1996).

Kourti, T, Lee, J. and MacGregor, J.F, Experiences with industrial applications of projection methods for MSPC. *Computer and Chemical Engineering*, **20(1996)**. 745-748. (1996)

Lam, H.L and Kamarul A.I, Improved MSPC for Chemical Process Fault Detection and Diagnosis via Outline Analysis Approach. 2nd World Engineering Congress, Sarawak, 39 – 44. (2002)

Lennox, B., Hiden, H.G, Montague, G.A., Kornfeld, G., Goulding, P.R., Application of MSPCto batch operations. *Computer and Chemical Engineering*, **Vol.24 (2000)** .291-296. (2000)

Mason, R.L, Tracy, N.D, and Young, J.C., Decomposition of T^2 for Multivariate Control Chart Interpretation. *Journal of Quality Technology*, **Vol. 27**. 99-108. (1995)

Nash, J.C. and Lefkovitch, L.P .Principal component and regression by SVD on a small computer.Apply Statistics (1976) **25, No. 3**, 210-216. (1976)

Westerterp K.R. and Ptasinsky K.J., Safety design of Cooled Tubular Reactions for Exothermic, Multiple Reaction; Parallel Reactions-I: Development of Criteria, *Chemical Engineering Science*, **vol. 39, No. 2**, 235-244. . (1984)